# Wikipedia コーパスに基づく”ユーモア”ある なぞかけ文生成システムの構築

山下健人 寺井あすか
公立はこだて未来大学 システム情報科学部

{b1013041, aterai}@fun.ac.jp

## 1 はじめに

言葉遊びの一つに「なぞかけ」というものがある.なぞかけとは与えられたお題 A に対し,B,C,C'の 3 つの言葉を用いて「A とかけて B ととくその心は C/C'」という形式の文を考える遊びである.前半の「A とかけて B ととく.」の部分で一見関係のない A と B という言葉を挙げ,後半の「その心は C/C'」の部分で C/C'という言葉を介してその何らかの共通点を示す[1].

本研究では,「パンとかけまして報道ととくその心はどちらも生地/記事がある」のような,A と関連する C と B と関連する C'の音韻の共通性により面白さを生じさせるなぞかけに焦点を当てる.上の例では,「パン」と「生地」,「報道」と「記事」は意味的に関連があり,「生地」と「記事」は音韻的に共通である.「パン」と「報道」という一見関係のない言葉を「生地/記事」という音韻により共通性を示している.このような同音の名詞C/C'をオチとしたなぞかけを対象とし,”ユーモア”あるなぞかけ文の生成システムを構築する.

ユーモアに関する理論として,意外な結びつきの発見によりユーモアが生起されると説明する不適合解消理論が存在する[2].この理論では,異種の組み合わせによる不適合がオチにより解消されることによりユーモアが引き起こされると説明される.さらに,実験的検討として,「A とかけて B ととくその心は C/C'」というなぞかけにおける不適合の解消と面白さの関与が定量的に示されている[3].すなわち,A,B に対しオチである C/C'の意外性(単語間の類似度)とユーモアの関連を示唆している.

さらに質問調査に基づくなぞかけ生成システムにより単語間の類似度とユーモアの関係が検討されている[1].先行システムはお題 A を入力することで A と C,B と C'両方の類似度が一定の範囲に含まれているものを対象として B,C,C'を抽出することでなぞかけを生成する.この A と C,B と C'の類似度の範囲は等しく設定されているが,なぞかけにおいて A と C,B と C'の類似度が同じ範囲にあるとは限らない.そのため,先行研究では,「A と C の類似度は高いが B と C'は類似度は低い」の様な類似度の範囲が異なる場合に生成されるなぞかけに関する検証が行われていない[1].また C と C'の類似度は考慮されておらず,C と C'の類似度の変化に伴うユーモアの評価も行われておらず,なぞかけのユーモアと単語間の類似度の十分な検証は行われていない.さらに,質問調査ではなぞかけの対象となる概念全てを網羅することは困難であり,質問調査に基づくシステムでは扱えるなぞかけの種類に限界がある.

そこで本研究ではWikipediaコーパスに基づき各単語間の類似度を推定することでより多くの概念を対象としたなぞかけ生成システムを構築する.さらにそのシミュレー

ション結果の評価によりユーモアのメカニズムとして単語間の類似度との関係を検証した.

## 2 なぞかけ生成システムの概要

なぞかけ生成システムは Wikipedia コーパス（6.8MB，見出し語数 1146584 語）に基づき,構築した[4].本システムにおける詳しいなぞかけ生成手順は以下の 9 ステップから成る.生成の流れを図 1 に示す.

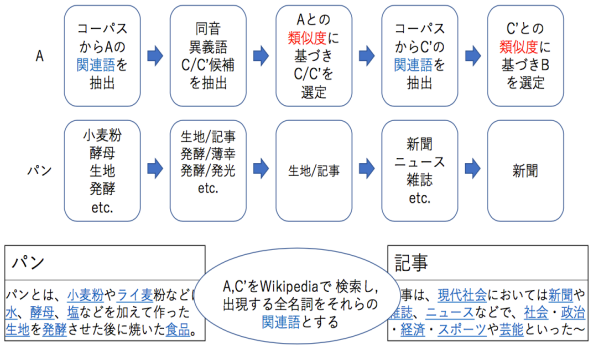


**図 1　システムの流れ**

1.入力されたお題 A を Wikipedia の見出し語から検索する.

2. Wikipedia の見出し語 A の本文に含まれる全名詞を抽出する.例えばパンをお題 A とした場合は小麦粉,ライ麦粉,水,酵母,生地などが抽出される

3. 2 で抽出した名詞からストップワードリストを用いて,Wikipedia の本文に頻繁に出現する「参照,全文,事,物」などの言葉を除外し,残った名詞を C の候補とする.ストップワードリストは SlothLib[5]に基づき設定した.

4. Wikipedia の見出し語から C の候補と同音語を探し,C と同音語の見出し語を C'の候補とする.

5. 4 より選定された候補 C と候補 C',A と C の類似度に基づき C/C'を選定する.

6. C'を Wikipedia の見出し語から検索する

7. Wikipedia の見出し語 C'の本文に含まれる全名詞を抽出する.

8. 7 で抽出した名詞からストップワードリストを用いて,Wikipedia の本文に頻繁に出現する「参照,全文,事,物,」などの言葉を除外し，残った名詞を B の候補とする.

9. その候補 B と C'の類似度に基づき,B を選定する.入力されたお題 A と上記ステップに選定された B,C,C'を用いて「A とかけて B ととくその心は C/C'です」というなぞかけ文の形で出力する.

## 3 シミュレーション

### 3.1 方法

Wikipedia コーパス内の名詞に対し word2vec を用いることで名詞ベクトルを推定し（word2vec のパラメータ：ベクトルの次元数 50,文脈の最大単語数 10,5 回未満登場する単語を削除）,cos 類似度を用いて名詞間の類似度を推定した.

今回,シミュレーションにおける類似度の範囲を高・中・低の 3 段階に設定した.A と C,B と C'の範囲を表 1 に,C と C'の範囲を表 2 に示す.表の中の sim(x,y)は x と y の類似度を表し,x と y は A と C,B と C',C と C'に該当する.A と C,B と C'の類似度の範囲は,シミュレーションに用いたお題全 10 種(パン,米,日本,地球,テレビ,テニス,パソコン,愛,神,雨)の Wikipedia 本文内の名詞と見出し語であるお題の類似度に基づき,それらの平均 0.2 と標準偏差 0.2 から平均±0.5 標準偏差を基準として設定した.C と C'の類似度の範囲は，Wikipedia コーパスからランダムに 5 千単語を選定し,それらの単語間の類似度に基づき設定した.ランダムに単語を選定した理由として,なぞかけの C と C'は同音異義語であり,特定の意味的な関係を持たない単語同士であると考えられるからである.平均類似度 0.16,標準偏差 0.2 に対し平均±0.5 標準偏差を基準とした.

**表 1　A と C,B と C'の類似度の範囲**

| | |
|---|---|
| 類似度 高 | sim(x,y)>0.3 |
| 類似度 中 | 0.1<sim(x,y)<0.3 |
| 類似度 低 | sim(x,y)<0.1 |

表 2　C と C'の類似度の範囲

| | |
|---|---|
| 類似度 高 | sim(x,y)>0.26 |
| 類似度 中 | 0.06<sim(x,y)<0.26 |
| 類似度 低 | sim(x,y)<0.06 |

上記の類似度の範囲を用いて,AとC,BとC',CとC',のそれぞれの類似度が高い場合,中程度の場合,低い場合の合計27パターンに対し,「パン,米,日本,地球,テレビ,テニス,パソコン,愛,神,雨」の10単語をお題Aとしてシミュレーションを行った.

### 3.2 結果

図 2 に 27 パターンのモデルで 10 個のお題をシミュレーションした際の,なぞかけ生成数を示す.A/C の類似度が低,B/C'の類似度が高,C/C'の類似度が低の場合に一番多くなぞかけが生成された.

また,パン(A/C 低 B/C'中 C/C'中),愛(A/C 低 B/C'高 C/C'高),雨(A/C 低 B/C'中 C/C'高),神(A/C 低 B/C'低 C/C'中),米(A/C 高 B/C'高 C/C'高),日本(A/C 高 B/C' 高 C/C'高),地球(A/C 低 B/C'高 C/C'高,A-C 低 B/C'低 C/C'中)の 8 つの場合になぞかけが 1 つも生成されなかった.

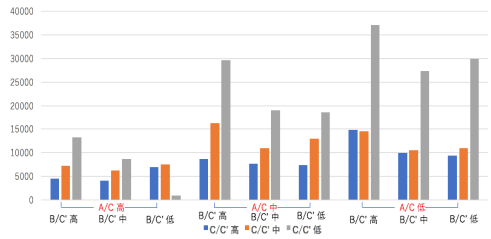


図 2 生成数

## 4 評価

### 4.1 方法

各類似度のパターンに対して生成されたなぞかけを用い,ユーモアとしてそれぞれの面白さに関する評価実験を行った.評価者は 18~24 歳の大学生,大学院生の男女 27 名である.システムが生成したなぞかけ 262 個(お題 10×27 パターン,なぞかけが生成されない場合 8 種を含む)を 5:面白い-1:面白くないの 5 段階で評価させた.

### 4.2 結果

得られた評価をもとに混合効果モデルを利用し分析を行った.その結果,1%水準で A/C の類似度の主効果: (F(2,32.6)= 16.771 ), B/C'の類似度の主効果: (F(2,31.0)= 26.775), A/C,B/C'の交互作用:(F(4, 4777.1)= 52.611 ),B/C',C/C'の交互作用: (F(4, 4777.3)= 12.065) が見られた.それらに関する多重比較の結果を図 3~6 に示す.

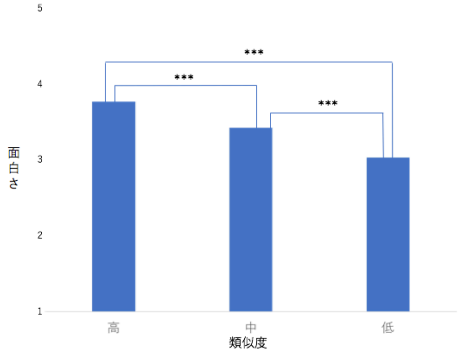


図 3　A/C の類似度の主効果(***p<.001)

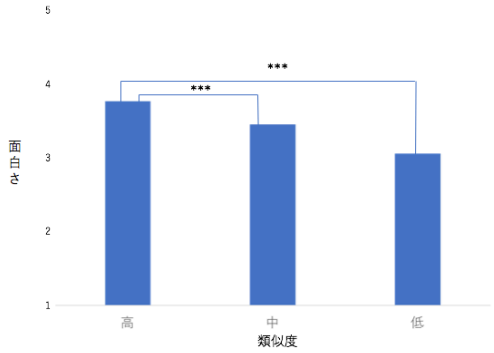


図 4　B/C'の類似度の主効果(***p<.001)

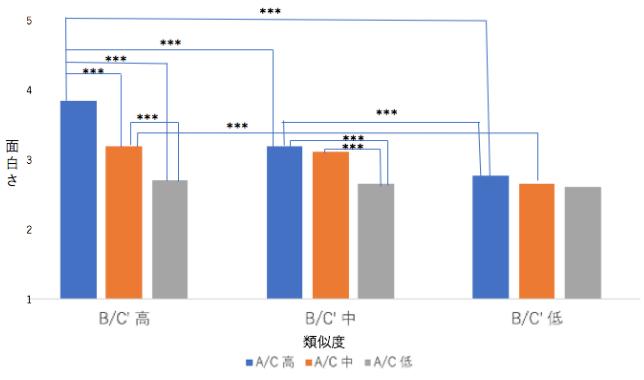


図 5　A/C,B/C'の交互作用(***p<.001)

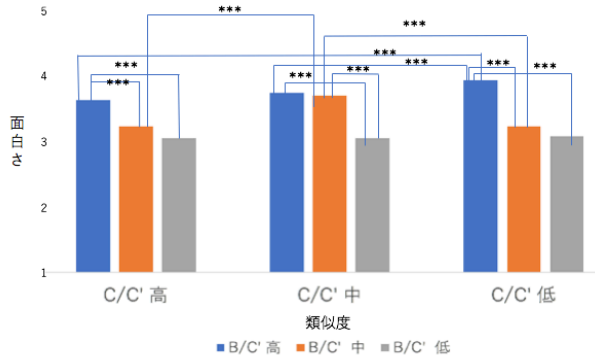


**図 6　B/C',C/C'の交互作用(***p<.001)**

A と C の単語間の類似度が高いほど,なぞかけの面白さが高く評価された(図 3).また,B と C'も同様に単語間の類似度が高いほど, 面白さが高い(図 4).B と C'の類似度,A と C の類似度が高くなるほど,なぞかけの評価が高くなる(図 5) (例:パンとかけまして書籍ととくその心はどちらも発酵/発行が関係します(A/C 高 B/C'高 C/C'中)).また,B と C'の類似度が高い場合,C と C'の類似度が低において高・中に比較し有意に評価が高く(例:雨とかけまして新約ととくその心はどちらも晴天/聖典が関係します(A/C 高 B/C'高 C/C'低)),B と C'の類似度が中程度の場合,C と C'の類似度が高,低に比較し中程度の場合に評価が高くなることが示された(図 6)(例:神とかけまして明治ととくその心はどちらも精霊/政令が関係します(A/C 高 B/C'中 C/C'低)).

## 5 考察

BとC'の類似度,AとCの類似度が高くなるほど,なぞかけの評価が高くなる,すなわちAとC, BとC'の単語の類似度が高く関連が理解しやすい場合に,面白さ（ユーモア）が生じることが示された.一方,BとC'の類似度が高い場合,CとC'の類似度が低い場合になぞかけの面白さが高くなる, BとC'の関連が理解しやすい場合はオチの同音異義語(C/C')の関係が遠く意外性を持つ場合に,面白さ(ユーモア)が生じる. しかし,Bと Cの類似度が中程度の場合,CとC'の類似度が中程度の場合に評価が高くなることから,Bと Cの関連性が多少理解しにくい場合は,CとC'の関連性もある程度理解しやすいことが必要であることが示された.

実験後に評価者から得られた感想として「見たことのないわからない単語が含まれたなぞかけがある」というものがあった.今回のシミュレーション・評価実験では単語の難易度に関する統制を取っておらず,類似度とユーモアの関連をより詳細に検討するには,普段あまり利用されない,理解しにくい単語をなぞかけ生成の対象から除外する必要があると考えられる.さらに現状のなぞかけ生成システムでは,「パンとかけて報道ととくその心は生地/記事です」のように C/C'の述部が「です」となるなぞかけが生成される.しかし,上記のなぞかけならば「パンとかけて報道ととくその心は生地/記事が重要です」のようにより適切な C/C'の述部を付加する事ができれば違和感のないより自然なぞかけになると考えられる.今後は「A とかけて B ととくその心は C/C'は D」といった A と C,B と C'の共通する関係 D を合わせて推定可能なシステムの構築が必要である.

## 参考文献

[1] 前田実香: 言葉の関連性とおもしろさを取り入れたなぞかけ生成, エンタテイメント感性特集, 5(3), 17-22, 2005

[2] Shultz, T. R., The role of incongruity and resolution in children's appreciation of cartoon humor. J. of Experimental Child Psychology, 13, 456-477. 1972

[3] 中村太戯留: 隠喩的表現において"面白さ"を感じるメカニズム,心理学研究 2009 年 第 80 巻 第 1 号, 2009

[4] mySoft「Wikipedia のダウンロード可能なデータ 」<http://www.mwsoft.jp/programming/munou/wikipedia_data_list.html>2018 年 1 月 15 日アクセス

[5] SDN 「slothlib」<https://ja.osdn.net/projects/slothlib/> 2018 年 1 月 15 日アクセス